DOCUMENT RESUME

ED 413 739 FL 024 760

AUTHOR Andersen, Poul
TITLE Cooperation with Central and Eastern Europe in Language Engineering.
PUB DATE 1995-00-00
NOTE 13p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Business Administration; *Computational Linguistics; Computer Software; *Computer Software Development; *Dictionaries; English (Second Language); Foreign Countries; Information Sources; *International Cooperation; *Language Planning; *Language Research; Pronunciation; Second Language Instruction; Telecommunications; Uncommonly Taught Languages; Vocabulary; Written Language
IDENTIFIERS Europe (Central); Europe (East); European Union

ABSTRACT

This paper outlines trends and activities in Central and Eastern European language research and language-related software development (language engineering) and briefly describes some specific projects. The language engineering segment of the European Union's Fourth Framework Programme, intended to facilitate use of telematics applications and increase options for communication within and between European language through language processing methods, is sketched, focusing on proposals for research funding. Ten joint research projects, involving partners from at least three countries in Eastern and Western Europe, are then described. These projects address: creation of a pronunciation lexicon for the European Union, with city and town names, street names, family names, and product names in 11 languages; creation of large speech data collections for Bulgarian, Estonian, Hungarian, Polish, and Romanian; software standardization and analytical tool and corpus development for Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovenian; compilation of terminologies in four areas (economics/management, energy, environment, telecommunications); extension of the generic computerized dictionary model; construction of morphological dictionary software; bilingual electronic dictionaries and intelligent text alignment; creation of modular courseware for English-as-a-Second-Language instruction; application of natural language processing techniques to technology for computer-assisted language learning; and development of a multilingual, multifunctional information retrieval system. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 739

Cooperation with Central and Eastern Europe in Language Engineering

Poul Andersen

DG XIII/E/6
European Commission
L-2920 LUXEMBOURG
Tel.: +352 4301 34324
Fax: +352 4301 34655
E-mail : poul.andersen@eurokom.ie
or Poul.Andersen@lux.dg13.cec.be

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Norbert
Volz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL024760

BEST COPY AVAILABLE

1. Introduction

The right of all citizens to communicate and receive information in their own language is a basic principle of the European Union. In order to preserve a multi-lingual Europe, with at present 11 official languages in the 15 Member States of the European Union, the European Commission feels naturally called upon to take a strong interest in Language Engineering (LE) along at least two dimensions:

- As a multilingual institution, it is itself an important user of LE products: The European Commission's Translation Service and other services use and support the development of Machine Translation and Machine-Aided Translation, Terminological Data Banks, and other tools that can help the Commission, especially in the huge task of translating more than 1 million pages each year.
  The Commission's own use of LE products is, however, outside the scope of this presentation.

- The European Commission encourages and financially supports research and development in the area of Language Engineering, with respect to the principle of subsidiarity, which implies that support mainly is given to multilingual activities, involving cooperation between partners from several member states.

2. Fourth Framework Programme

At present, scientific and technological cooperation in the field of LE is supported through the *Fourth Framework Programme for Research and Technological Development (1994–1998)* (4th FWP). This programme comprises a wide range of areas with a total budget of approximately 12 300 million ECU. The largest part of the funding is allocated to *Activity 1*, which covers RTD (Research and Technological Development) and Demonstration Programmes within the European Union. One of the Specific Programmes within Activity 1 of 4th FWP is the *Telematics Applications Programme*, which includes *Language Engineering* with a budget of approximately 81 million ECU.

The aim of Language Engineering is to facilitate the use of telematics applications and to increase the possibilities for communication in and between European languages by integrating new spoken and written language-

processing methods. Work focuses on pilot projects that integrate language technologies into information and communications systems and services. A key objective is to improve their ease of use and functionality and broaden their scope across different languages.

Language Engineering as defined in 4th FWP covers the following *Action Lines*:

1. **Pilot Applications**
   Document Creation and Management
   Information and Communication Services
   Translation and Foreign Language Acquisition
2. **Re-usable Language Resources**
3. **Language Engineering Research**
4. **Support Issues Specific to Language Engineering**
   (i.e. standards, assessment and evaluation, awareness activities, user surveys)

Since 1995, the Telematics Applications Programme, including Language Engineering, has been open to participation of institutions from Central and Eastern Europe, who can receive funding from the budget allocated to Activity 2 (see below).

The Third Call for Proposals under this programme was published on September 15, 1995, with a closing date on January 15, 1996. A final call is planned for publication on September 15, 1996.

A home page for Language Engineering can be found at:

HTTP://www.echo.lu/programmes/en/LangEng/le.html

All specific inquiries regarding Language Engineering within 4th FWP can be obtained from:

European Commission
DG XIII-E-5 LE Office
Batiment Jean Monnet (B4-002)
L-2920 Luxembourg
Fax: +352 4301 34999

2.1. International cooperation in the Fourth Framework Programme

A part of the Framework Programme of special interest to institutions in countries outside the European Union, is Activity 2, which covers Cooperation with *third countries and international organisations* with a total budget of 540 million ECU. Almost half of this budget is used to support cooperation with *Central and Eastern Europe.*

The main goals for RTD cooperation with Central and Eastern Europe are:
- to help to safeguard the RTD potential in these countries;
- to help to solve important social, economic, and ecological problems;
- to intensify cooperation in RTD fields where these countries are in the forefront on a world level.

The precise list of countries from Central and Eastern Europe that are eligible for support may change with the political development, e.g., in ex-Yugoslavia. In the latest two calls for proposals 1994 and 1995/96, the following groups of countries could participate:
- Countries of Central Europe (CCE):
  Estonia, Latvia, Lithuania, Poland, Czech Republic, Slovakia, Hungary, Slovenia, Romania, Bulgaria, Albania.
- Newly Independent States (NIS):
  Russia, Belarus, Ukraine, Moldova, Armenia, Georgia, Azerbaidjan, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan.

The normal procedure for allocating funds is through *Open Calls for Proposals*.

A Call for Proposals was published on *October 17, 1995* with a deadline for receipt of proposals on *February 29, 1996*. A second Call for Proposals is planned for publication on *March 15, 1997*. These Calls are published in the EC Official Journal.

In order to participate in a Call for Proposals, it is necessary not only to read the text of the call in the Official Journal, but also to obtain the *information package* for the call, which contains detailed and specific information about what kind of proposals are eligible, as well as application forms to fill in with scientific, financial, and other administrative information. Such information packages are sent out to potential proposers, already known to the European Commission, and to anyone who asks for them.

The information package for the latest call with closing date on September 29, 1996 invites proposals for projects within Language Engineering, which ...
... must contribute to an open language infrastructure between the EU and CEC/NIS by focusing on:
- *creation of new language resources for the CEC/NIS languages;*
- *augmentation and further development of existing language resources for these languages;*
- *validation and exploitation of such resources for later integration into a range of computer-based services and products such as document management and translation tools which specifically identified needs in CEC/NIS.*

It also invites proposals for *Support Actions*, such as Awareness Seminars (see below).

Information packages are not always available in printed form at the date of publication of a call, but it is possible to collect this information over the INTERNET from

http://WWW.cordis.lu/

which contains an *Electronic Document Delivery Service* for documents and other texts related to the Fourth Framework Programme.

3. Implementation

Most of the funds under 4th FWP are spent on 3 types of activities:
- Concerted Actions
- Joint Research Projects
- Accompanying measures

3.1 Concerted Actions

In order to promote the creation of networks of scientists in the public and private sectors, the Commission supports *Concerted Actions*, which bring together teams from Eastern and Western Europe. Such Concerted Actions make it possible to establish permanent cooperation links which can serve as a basis for all kinds of research activity, and they encourage interactions between various disciplines, transfer of technologies, dissemination of results, and exchange of information in general. They encourage cooperation between academies and industries, help to identify new partners, and put research workers in contact with each other and with the responsible authorities in the different countries.

The intervention of the Commission covers coordination expenses: meetings, workshops, distribution of information, and exchange with and visits to other institutions taking part in the action. Financing can also be given for centralised facilities such as data banks, specialised communication facilities, and preparation and distribution of reference materials. They normally do not include funding of research, which the participating institutions are expected to get funded by other means, e.g., through EU-funded Joint Research Projects (see below).

By their nature, Concerted Actions are typically 'flat' structures with many participants from different countries.

The last Call for Proposals ***(COPERNICUS 1994)*** was published on January 31, 1994 with a deadline for receipt of proposals on April 29, 1994, and resulted in the funding of two Concerted Actions involving Central and Eastern Europe in the area of Language Engineering, both with Romanian participation:

- **TELRI – Trans-European Language Resources Infrastructure**
  TELRI has participants from all 11 *Countries of Central Europe*, as listed above. TELRI is coordinated by the *Institut für deutsche Sprache*, Mannheim, Germany, represented by Dr. Wolfgang Teubert. The other Western European participants represent leading institutions in Great Britain, France, Italy, the Netherlands, and Sweden, most of which are involved in networking activities within Western Europe such as ELRA *(European Language Resources Association)*, and can thus serve as a link between such activities and Central European institutions.
- **ELSNET goes East**
  ELSNET goes East is an extension of ELSNET *(European Network in Language and Speech)* to Central and Eastern Europe. Its geographical coverage differs from TELRI, as it does not have participants from all 11 Central European countries; however, it has participants from Russia and Belarus.

3.2 Joint Research Projects

These projects aim to assemble, for a specific research subject, a multinational team to perform research and development work and to obtain results in collaboration. A Joint Research Project typically comprises 3–6 partners coming from at least 3 different countries in Eastern and Western Europe. The duration of a project is normally between one and three years.

The above mentioned Call for Proposals, ***COPERNICUS 1994***, resulted in the funding of the following 10 Joint Research Projects involving Central and Eastern Europe in the area of Language Engineering.

3.2.1. 'Broad' projects with many partners
→ good potential for creating infrastructure and networking in specific areas

Two Projects Within Speech:

- **ONOMASTICA-COPERNICUS**
  ONOMASTICA builds a pronunciation lexicon for the European Union, with city and town names, street names, family names, product names in 11 languages – Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish, and Swedish.
  ONOMASTICA-COPERNICUS extends the languages covered in ONOMASTICA to include names and pronunciations for *Czech, Estonian, Latvian, Polish, Romanian, Slovakian, Slovenian, and Ukrainian.*
  Pronunciation dictionaries for up to 250 000 names per language will be constructed, and quality controlled pronunciation lexicons will be made available in machine readable form (CD-ROM) for use in automated language systems by international European companies in the telecommunications sector and in the European (dictionary) publishing industry, as well as by language system researchers and developers.

- **BABEL – A Multi-Language Database**
  BABEL pursues the creation of large speech data collections for *Bulgarian, Estonian, Hungarian, Polish and Romanian:*
  - a speech database containing both read (about 75%) and spontaneous (about 15%) utterances collected from 100 speakers;
  - some spontaneous utterances will be accompanied by subsequent read versions (about 10%). It is planned to produce one CD-ROM disk per language (around 6 hours of speech).
  - a text corpus containing the orthographic text of the read utterances for each speaker.

  The project aims to produce phonetically and prosodically labelled annotations of at least 15% of the recorded material, using as far as possible semi-automatic labelling techniques, but ensuring expert checking.

One Project within Corpus Linguistics:

- **MULTEXT-EAST**
  MULTEXT-EAST is a spin-off of one of the largest EU projects in the domain of language tools and resources, MULTEXT, which had three main objectives:

- *Standardization*: development of a software standard based on a "software Lego" approach for corpus handling tools, together with TEI-based encoding conventions specifically suited to multilingual corpora and language engineering applications.
- *Tool and corpus development*: development of an extensive set of tools for corpus annotation and exploitation as well as the first annotated large-scale multilingual corpus for EU languages, intended to serve as a reference and test-bed for multilingual tools and applications.
- *Industrial validation*: integration by six major European companies of project results into high-level NLP applications such as term extraction and machine translation lexicon generation, thus providing a first indication of downstream applicability.

MULTEXT-EAST extends MULTEXT by transferring its expertise, methodologies, and tools to CEE countries, and the two projects together create a network of more than twenty academic research centers and companies, developing and using common lingware and methodologies, as well as producing the first annotated large-scale multilingual corpus for 12 EU and CEE languages.

East European languages covered: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovenian.

One project within Terminology:

- **PRACTEAST - Preparatory Actions for Terminological Assistance to Central and Eastern European Countries**

The concrete outcome of PRACTEAST is the compilation of four terminological collections in the domains of Economics and Management, Energy, Environment, and Telecommunications, each collection containing the 2000 most common terms with English terms and definitions and French and Spanish equivalent terms.

The databases compiled by the Coordinating Contractor, will revert to the CEC or to the infrastructure created as a result of the tasks completed within the framework of other EU funded projects. For example, the EURODICAUTOM database could be enriched with no less than 11 new European languages.

Furthermore, each partner will be able to use the Multilingual Database for its own research purposes and to publish the corresponding conventional paper dictionaries.

East European languages covered: Bulgarian, Czech, Estonian, Hungarian, Latvian, Lithuanian, Polish, Romanian, Russian, Slovakian, and Ukrainian.

3.2.2 'Narrow' projects with few partners
→ good potential for creating resources and/or applications in specific areas

Three projects concerned with Dictionary standards + coding:

- **CEGLEX - Central European GeneLEX model**
  CEGLEX aims at extending the generic electronic dictionary model (and accompanying SGML DTD) developed in the EUREKA GENELEX project to three central European languages and to:
  - give rapid access to a Western European linguistic engineering pre-standard model for Central European actors of the NLP scene,
  - extend the GENELEX model to new languages, evaluating its appropriateness and making it a stronger candidate for being an internationally recognised standard,
  - start a larger cooperation between the Partners leading to industrial level applications.

  The work will consist in
  - identifying theoretical issues in *Czech, Hungarian, and Polish* that may lead to specific adaptations and extensions of the model;
  - building a representative core lexicon that will conform to the elaborated model;
  - verifying the possibility to use the core dictionary in the context of an application.

- **GRAMLEX**
  The aim of GRAMLEX is to facilitate the initiation, coordination, and standardisation of the construction of morphological dictionary packages for French, *Hungarian,* Italian, and *Polish,* including detailed formal description of the morphology of the languages. The major challenges in such an enterprise are to give the description the largest possible coverage in order to be able to process unrestricted text; to share as many as possible of the formats, methods, and algorithms; and to improve time and space efficiency of programs.
  Keywords: lexical tagging, morphological dictionaries, lexical resources.

- **BILEDITA – Bilingual electronic dictionaries and intelligent text alignment**
  The goals of BILEDITA are
  - to provide a uniform dictionary format for all of the existing dictionaries constructed by the partners;
  - to provide a uniform lexical encoding scheme for both form and con-

tent of the entries in the electronic dictionaries. This has been systematically dealt with as far as the French and German dictionaries are concerned, and will be accomplished for the other project languages: *Bulgarian, Polish, and Russian;*
- to systematically construct and exploit bilingual corpora for the purpose of building bilingual basic dictionaries, terminology dictionaries, and phrasal dictionaries.

Two projects within CALL (Computer Assisted Language Learning):

- **BALTIC – Basic and Advanced Language Transnational Interactive Course**
The objective is to create modular courseware for computer-assisted teaching of English to the citizens of *Latvia, Estonia, and Lithuania,* that will allow self-teaching, classroom teaching, and long distance teaching via an on-line network. BALTIC uses an already existing Italian/English course as a base, implementing the parts that allow the passage to other languages.

- **GLOSSER**
GLOSSER applies NLP techniques, especially morphological processing and corpora analysis, to technology for computer-assisted language learning (CALL) with potential spin-offs in translation technology, information retrieval, and text-indexing technologies.
GLOSSER's aim is to enable speakers of *Bulgarian, Hungarian, or Estonian,* who are intermediate language learners/users of English, to read and learn English more fluently. For example, when he reads a software manual on the screen and encounters an unknown word or an unfamiliar use of known word, he can point to it with the mouse and invoke online help, which will provide him with the following facilities:
  - a morphological parse, separating stem and ending, together with an explanation of the significance of the inflection;
  - the entry to the word in a bilingual X/English or a monolingual English dictionary;
  - (for a small number of words) an audible pronunciation;
  - access to similar examples of the word in online bilingual corpora.

One Speech project:

- **SQEL – Spoken Queries in European Languages**
SQEL aims at the development of a multi-lingual and multi-functional information retrieval system, based on an existing, experimental infor-

mation dialogue system for English, German, French, and Spanish, SUNDIAL. Within SQEL, a prototype of such a system will be developed for *Czech, Slovak, and Slovenian.*

3.3 More Information on Concerted Actions and Joint Research Projects

A common home page for all 10 Joint Research Projects and the 2 Concerted Actions from the 1994 Call for Proposals can be found on the INTERNET, under

http://www.fwi.uva.nl/research/illc/ege/cop.le.proj.html

3.4 Accompanying Measures and Support Actions

Preparatory, accompanying, and support measures comprise i.a. the following activities:

- **National or Regional Awareness Seminars** are being conducted in Riga (for the three Baltic states), Prague, Poznan, Bucharest, and St. Petersburg in 1994-1996, and more are planned for other Central and Eastern European countries and regions. These seminars are aimed at opinion formers, media, providers, the research/academic community, users, and government organisations. Approaching these groups to make them realise the benefits of undertaking initiatives in the language engineering field, and the risks of not doing it, are some of the key subjects.
- **Information gathering** is supported, in order to identify possible cooperation partners in the area of language engineering in Central and Eastern Europe, and subsequently information dissemination to these partners. This information gathering is an important side-effect of the Awareness Seminars, and the Commission's recent, more systematic strategy towards the integration of Central and Eastern Europe into Trans-European cooperation activities started with a small seminar, which took place in Luxembourg in January 1994 and resulted in a first overview over the state of affairs in most of the countries.
  The seminar in Luxembourg was also used for the launch of a study carried out by ELSNET for the European Commission; the results of which were published in October 1994 as *Survey of Language Engineering Organisations in Central and Eastern Europe*. This document contains profiles of over 100 language engineering organisations in the following Central and Eastern

Europe and Newly Independent States: Belarus, Bulgaria, Czech Republic, Estonia, Georgia, Hungary, Latvia,Lithuania, Poland, Romania, Russia, Slovakia, Slovenia, Ukraine.

The document is available over the INTERNET at

http://www.cogsci.ed.ac.uk/elsnet/survey/survey.html

This survey is obviously not complete, but the information gathering is continued through the abovementioned Concerted Actions TELRI and ELSNET goes East, and will eventually result in an updated and more detailed picture.

4. More Information ...

about Calls for Proposals under Activity 2 of 4th FWP can be found

- on the Web at *http://www.cordis.lu* (as mentioned above);
- from local contact persons in each of the 11 eligible Central European countries and in Belarus, Russian Federation, Ukraine, and Georgia;
- directly from the European Commission:

Grazyna WOJCIESZKO
DG XIII
Tel.: +32-2-295 83 57
Fax: +32-2-296 17 16
E-mail : gwoj@dg13.cec.be

Inquiries about Trans-European cooperation activities related to language engineering may also be addressed to the author of this presentation.

An electronic mailing list *Eastern (Europe) Language Engineering* comprises more than 100 persons from Western as well as Central and Eastern Europe with a special interest in Trans-European cooperation. This mailing list can be used for announcement of conferences and other events, and any member of the list can enter search for cooperation partners in specific projects etc.

In order subscribe to this list, please send an E-mail to

poul.andersen@eurokom.ie